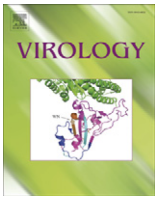
Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro



Evolutionary dynamics of *Newcastle disease virus*[☆]

Patti J. Miller[a], L. Mia Kim[a,1], Hon S. Ip[b], Claudio L. Afonso[a,*]

[a] *Southeast Poultry Research Laboratories, USDA ARS, Southeast Poultry Research Laboratory, 934 College Station Rd., Athens, GA 30605, USA*
[b] *USGS National Wildlife Health Center, 6006 Schroeder Road, Madison, WI 53711-6223, USA*

ARTICLE INFO

Article history:
Received 23 February 2009
Returned to author for revision
27 March 2009
Accepted 22 May 2009
Available online 28 June 2009

Keywords:
Newcastle disease virus
Evolution
Recombination
Selection
Virulence
APMV-1

ABSTRACT

A comprehensive dataset of NDV genome sequences was evaluated using bioinformatics to characterize the evolutionary forces affecting NDV genomes. Despite evidence of recombination in most genes, only one event in the fusion gene of genotype V viruses produced evolutionarily viable progenies. The codon-associated rate of change for the six NDV proteins revealed that the highest rate of change occurred at the fusion protein. All proteins were under strong purifying (negative) selection; the fusion protein displayed the highest number of amino acids under positive selection. Regardless of the phylogenetic grouping or the level of virulence, the cleavage site motif was highly conserved implying that mutations at this site that result in changes of virulence may not be favored. The coding sequence of the fusion gene and the genomes of viruses from wild birds displayed higher yearly rates of change in virulent viruses than in viruses of low virulence, suggesting that an increase in virulence may accelerate the rate of NDV evolution.

Introduction

Newcastle disease virus (NDV), synonymous with avian paramyxovirus-1 (APMV-1), is a member of the *Avulavirus* genus in the *Paramyxoviridae* family. Encompassing a diverse group of single-stranded, negative sense, non-segmented, enveloped RNA viruses of approximately 15.2 kb, NDV have a broad host range and are known to infect over 200 bird species (Alexander et al., 2003). The NDV genome encodes for six major structural proteins: nucleocapsid (NC), phosphoprotein (P), matrix (M), fusion (F), hemagglutinin-neuraminidase (HN), the RNA dependent RNA polymerase (L), and also for a seventh protein (V) that is produced by a frame shift within the P coding region.

Viruses of low virulence are often utilized as vaccines and typically cause asymptomatic infections or mild respiratory disease. Newcastle disease (ND) results from infections with virulent NDV, defined as strains having intracerebral pathogenicity indices (ICPI) of ≥ 0.7 in day old chickens (*Gallus gallus*) and/or having multiple basic amino acids (at least three arginine (R) or lysine (K) residues) at the C-terminus of the fusion protein cleavage site along with a phenylalanine at position 117 (OIE, 2004). Clinical signs of ND range from moderate respiratory disease with occasional nervous signs to acute death. The more severe forms of the disease are categorized by the organ system which is affected: the viscerotropic form causes extensive hemorrhage in multiple gastrointestinal organs with mild nervous signs, and the neurotropic form predominantly affects the central nervous system with no other gross lesions (Alexander et al., 2003).

Low virulence NDV have monobasic fusion cleavage site motifs at amino acid (aa) positions 112–113 and 115–116 and a leucine (L) at position 117 of the F protein (Glickman et al., 1988), and can only be cleaved by trypsin-like enzymes found in the respiratory and intestinal tracts, which restricts their replication to these systems (Rott, 1979). In contrast, virulent viruses have multiple basic amino acids in the fusion cleavage site: 112R-K/R-Q-K/R-R-F117. As few as two nucleotide changes can result in emergence of a virulent form of NDV from a low virulence virus; however, there are only a few documented cases of this occurrence. The outbreaks that occurred in Ireland in 1990 and in Australia from 1998 through 2000 were each the result of low virulence viruses mutating to high virulence. In Ireland, the low virulence viruses were endemic in coastal wildlife populations, and in Australia, low virulence viruses initially circulated in poultry (Alexander et al., 1992; Gould et al., 2001).

Evidence suggesting that low virulence NDV can become highly pathogenic in poultry has spurred considerable interest in understanding the evolutionary forces affecting genomic changes. Despite this interest, few studies provide insights into the evolutionary mechanisms affecting genomic changes. The majority of genomic changes in non-segmented RNA viruses are due to either the intrinsic error rate of the polymerase or as a result of recombination. Polymerase error generates a large number of genetic variants, known as quasispecies, upon which natural forces select changes in

[☆] Use of trade or product names does not imply endorsement by the U.S. Government.
[*] Corresponding author. Fax: +1 706 546 3161.
E-mail address: claudio.afonso@ars.usda.gov (C.L. Afonso).
[1] Current address: Animal Health Service, Food and Agriculture Organization of the United Nations, Viale delle Terme di Caracalla, 00153 Rome, Italy.

0042-6822/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2009.05.033

the NDV genome (Duarte et al., 1994). The presence of selection pressures at specific amino acid sites within proteins is recognized as adaptive evolution. Positively selected sites indicate that past evolutionary pressures led to increased genetic variation, while negative (purifying) selection reflects the tendency toward sequence conservation (Bush, 2001; Kosiol et al., 2006; Yang and Nielsen, 2002). Recombination, although common among positive-stranded RNA viruses that encode their own RNA polymerase (Lai, 1992; Worobey and Holmes, 1999), is infrequently reported in the non-segmented negative-strand RNA viruses (Chare et al., 2003; Spann et al., 2003). Widespread recombination events, identified in GenBank datasets, have sparked controversy regarding the role of recombination in NDV evolution (Afonso, 2008).

Understanding the role of virulence on the evolutionary dynamics of both virulent and non-virulent NDV can help predict and prevent future outbreaks. Here we analyzed the role of recombination, selection pressures, and virulence on the evolutionary changes of NDV proteins with emphasis on the F protein cleavage site that is responsible for virulence.

Results

One event compatible with the action of recombination is of evolutionary significance

Recombination analysis of complete coding regions in NDV isolates obtained through sequencing and from available GenBank datasets were performed using the split decomposition method and by six local statistical methods as implemented in the software program RDP 3.24 (data not shown). While statistically significant recombination events were identified for every gene except the polymerase (nucleocapsid/$n=10$, phosphoprotein/$n=16$, matrix/$n=14$, fusion/$n=14$, hemagglutinin-neuraminidase/$n=3$) (data not shown and Supplementary Table 2), only one breakpoint fulfilled the stringent criteria for predicting an evolutionary role for recombination.

This event, compatible with recombination, was identified in the *F* gene of an NDV isolate from a turkey (Fig. 1; 92US08TKY) with a beginning breakpoint at nucleotide position 103, using four different

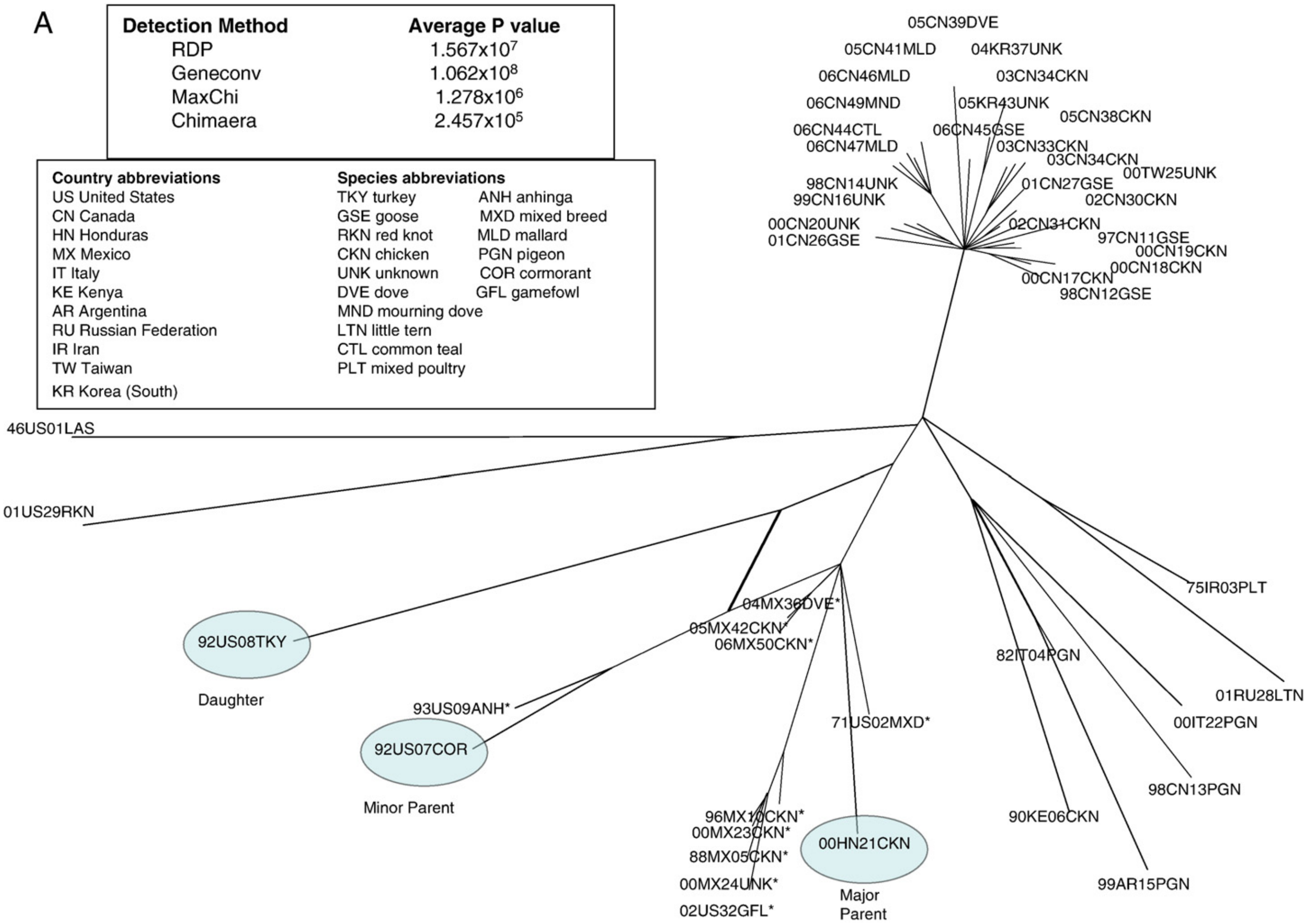


Fig. 1. (A) Detection of recombination on the complete fusion protein of virulent NDV. The split decomposition method, as implemented in SplitsTree 4, was used to represent a lineage of viruses from genotype V that descended from a common recombination event. In the graph, phylogenies with potential for recombination produce a tree-like network with parallel branches. Viruses marked with an asterisk (*), shared with 92US08TKY, 92US07COR as one parent and an undetermined parent. (B–D) Graphical representation of recombination support, for daughter 92US08TKY, using the programs Bootscan, Geneconv and RDP, respectively. Results of analysis of the NDV genome provide evidence compatible with the action of RNA recombination at position 103 (▲) detected by these three methods. The *X*-axis in B–D shows the nucleotide position of the NDV genome. The *Y*-axis gives % Bootstrap support (100 replicates) for the Bootscanning method (B), Log $_{10}$ [KA *P*-value] for the Geneconv program (C), and pairwise identity for the RDP method (D). Maximum-likelihood phylogenetic tree reconstruction at either side of the putative break point of the fusion protein of viruses of genotype V; (E) region between nucleotide 1 and 103; (F) Region between nucleotide 104 and 1653. The putative recombinants detected are outlined with boxes. Virus designations represent an 8 to 10 character name containing the two-digit year of collection, location of abbreviation, unique virus identification (one to three characters), and species abbreviation. Line colors designate the two sequences that are being compared in each program.

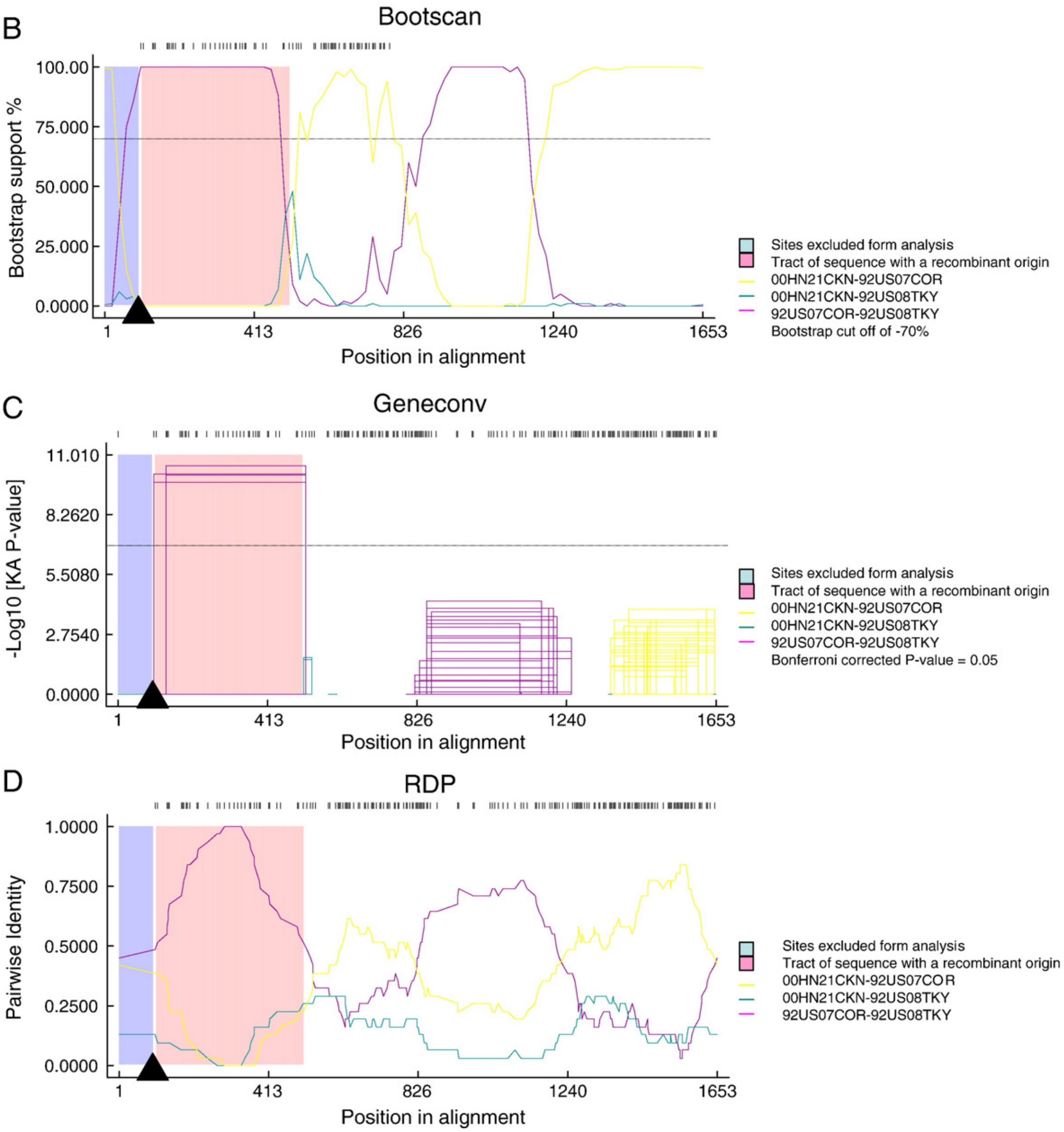


Fig. 1 (*continued*).

recombination detection methods (Figs. 1A–D). This daughter virus, 92US08TKY, appeared to arise from a recombination event between 92US07COR and 00HN21CKN with statistically significant probabilities ($P<0.01$) obtained with RDP, Geneconv, MaxChi and Chimaera (see inset of Figs. 1A and B–D). These ten viruses isolated from 1971 through 2006 in North America, which shared a virtually identical breaking point with recombinant sequences at position 97 detected using RDP (4.37×10^{-2}), Geneconv (1.24×10^{-4}), Bootscan (2.32×10^{-2}), MaxChi (1.29×10^{-2}), Chimaera (2.69×10^{-3}), SiScan (3.11×10^{-30}) and 3Seq (8.94×10^{-15}). The relationship among these recombinant viruses is illustrated with a tree-like network containing parallel branches in Fig. 1A, created using a split decomposition algorithm.

While these eight viruses shared the 92US07COR parent, their other predicted parent virus was undetermined by the analysis method used. To confirm the results obtained through RDP, additional maximum likelihood phylogenic trees were constructed with the nucleotide sequences from both sides of the breaking point (Figs. 1E and F). The putative recombinant daughter virus, 92US08TKY, groups and is closer in distance with parent virus, 00HN21CKN, prior to the break point (Fig. 1E) and with parent 92US07COR after the break point (Fig. 1F). Because most of these viruses were isolated independently at different time points and/or locations, this event compatible with the action of RNA recombination and shared by these genotype V viruses, appears to have occurred naturally and not as the result of a laboratory artifact (Fig. 1A). A second breakpoint is around nucleotide position 850 that is supported by Bootscan (Fig. 1B), but not by the Geneconv or RDP. This breakpoint was not shared by all of the viruses and it may not be a real event.

All six NDV complete coding regions display different rates of amino acid changes, widespread negative selection and positive selection evident only at a few select amino acid sites

To identify additional forces affecting NDV evolution first we determined the overall codon substitution rates and estimated average selective pressures per protein by using Model 0 of CodeML. This model assumes a constant substitution rate for all sites and provides estimated values averaged across all sites. The gene coding for the surface F protein was found to be the most variable (rate = 6.94 per codon), while the gene that encodes the nucleocapsid protein was

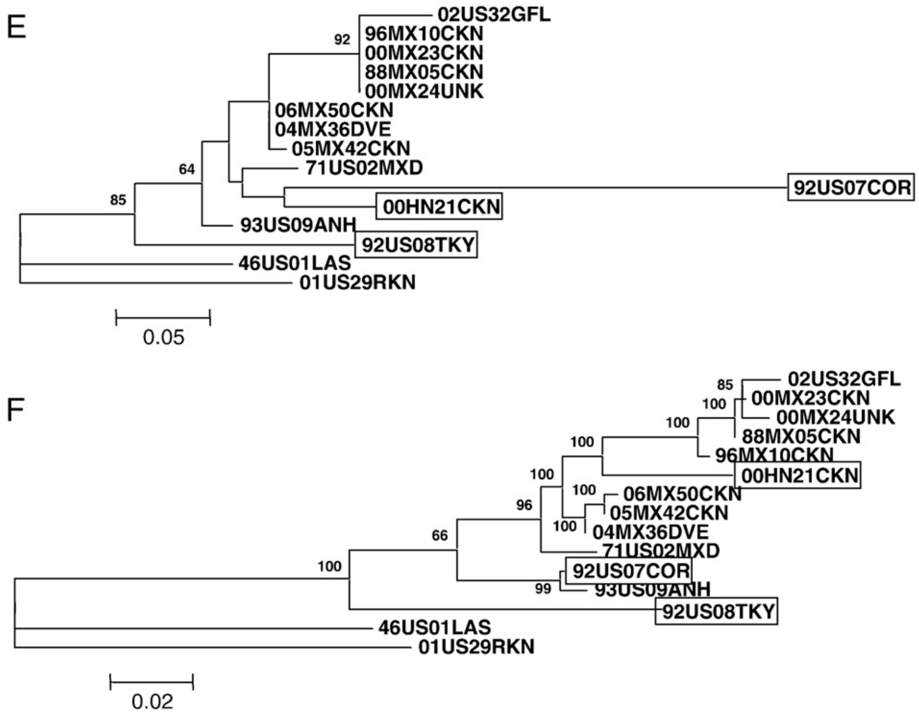


Fig. 1 (*continued*).

the most conserved (rate = 3.34). Rates for the remaining genes ranged from 4.37 to 6.1 (HN: 4.39), P: 4.37, M: 6.1, and L: 4.39). To determine if overall codon substitution rates were associated with adaptive evolution we compared the d*N* (non-synonymous substitutions per synonymous site) to d*S* (synonymous substitutions per synonymous site; Table 1). The interpretation of the d*N* to d*S* ratio is as follows: $\omega > 1$ is indicative of positive selection, $\omega = 1$ indicates neutral selection, and $\omega < 1$ indicates negative or purifying selection.

Table 1
Overall non-synonymous to synonymous ratios (ω) for all NDV proteins under Model 1 and positively selected codons (d*N*/d*S*>1) determined by Model 8 of CodeML.

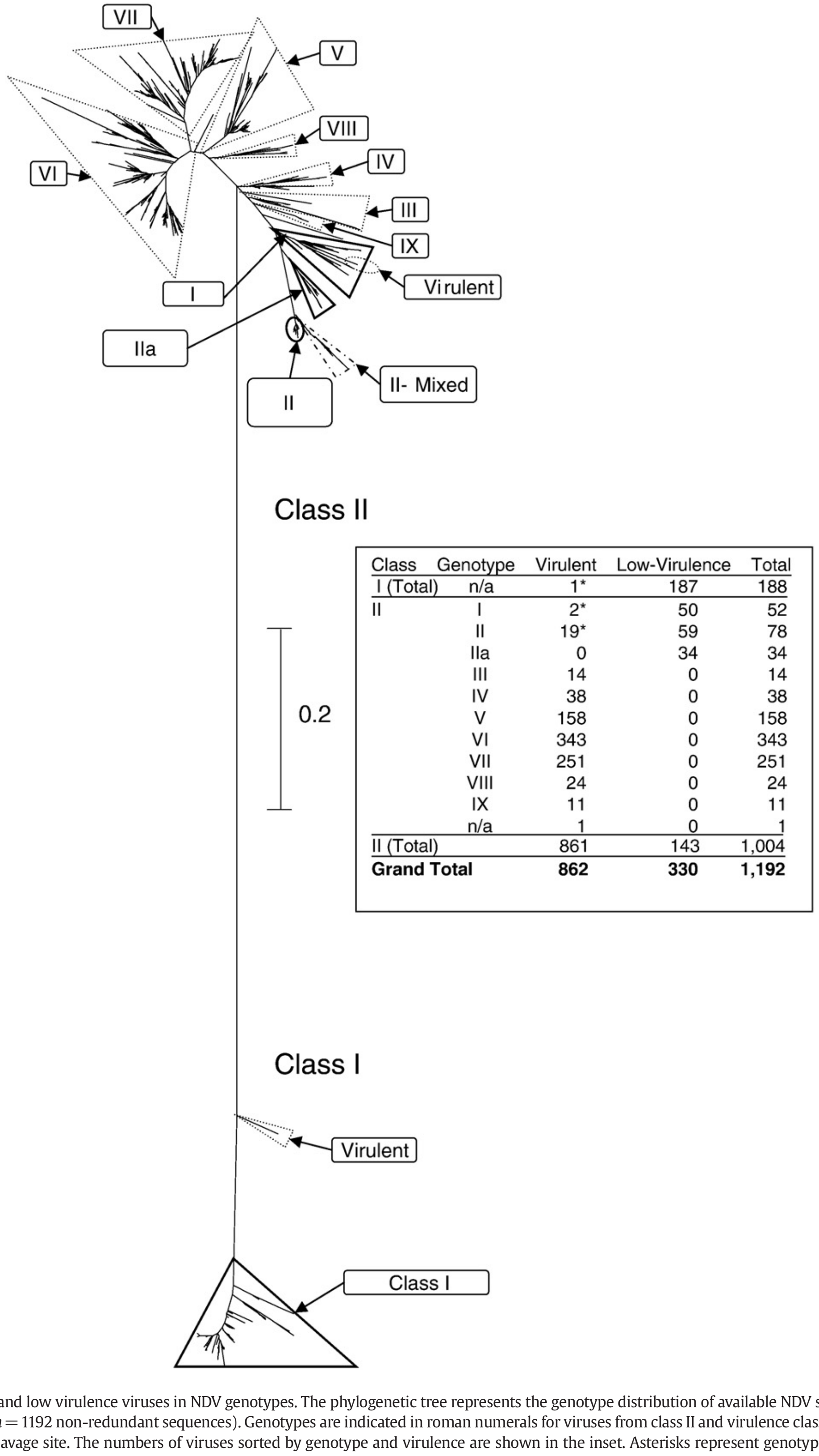
Gene	Recombinant and non-recombinant				Non-recombinant				
	n	Overall d*N*/d*S*	Variance	Sites with d*N*/d*S*>1	*n*	d*N*/d*S*	Overall variance	*P*	Sites with d*N*/d*S*>1
F	129	0.107	0.031	4	115	0.102	0.029	0.999	4
				5				0.999	5
				9				0.974	17
				10				0.994	19
				11				0.963	20
				17				0.985	22
				19				0.999	27
				20				0.999	28
				22					
				27					
				28					
				516					
HN	122	0.141	0.027	4	119	0.141	0.027	0.978	4
				266				0.905	266
				540					
M	74	0.103	0.114	259	70	0.097	0.013	0.977	259
NC	82	0.079	0.021	406	72	0.074	0.016	0.994	406
				431				0.913	430
				434					
				469					
P	91	0.257	0.060	72	75	0.254	0.058	None	
L	43	0.041	0.007	259	43	0.007	0.041	0.921	259
				417				0.997	417

F: fusion; *HN*: hemagglutinin-neuraminidase; *M*: matrix; *NC*: nucleocapsid; *P*: phosphoprotein; *L*: polymerase; *n* = number sequences analyzed; posterior cutoff for sites with d*N*/d*S*>1 = 0.9; *P* = posterior probability of sites with d*N*/d*S*>1.

Ratios were obtained for all six coding regions using two datasets: a complete dataset with evidence of recombination and one that excluded recombination (Supplementary Table 2). Interestingly, the highest overall ω values corresponded to the *P* gene ($\omega = 0.254$) that also encodes the V protein involved in suppressing the interferon response. Each of the remaining genes also displayed evidence of negative selection (ω (F) = 0.102, ω (HN) = 0.141, ω (M) = 0.097, ω (NC) = 0.074, and ω (L) = 0.007; Table 1) affecting only a small proportion of the overall changes occurring in NDV proteins. Not surprisingly, the *NC* and *L* genes displayed overall lower rates of change and higher levels of negative selection as they are involved in virus replication and therefore less likely to suffer selective pressure from the host immune system. The identification of the specific sites undergoing positive selection by CodeML Model 8 is presented in Table 1 in which both the complete and recombination-free datasets were included. Table 1 compiles the positions of positively selected amino acids across all six proteins. When only the proteins with no evidence of recombination were analyzed eight, two, one, two, zero, and two amino acids were found to be under positive selection for the F, HN, M, NC, P and L, respectively. However, when evaluating the complete GenBank dataset (including proteins with evidence of recombination) twelve, three, one, four, one, and two positively selected sites on each protein were produced for the F, HN, M, NC, P and L, respectively. While the number of positive selected sites was certainly affected by the presence of recombinant sequences in the dataset, positively selected sites could clearly be identified in the non-recombinant dataset. Positive selection was notably absent in the F protein at the critical cleavages site (amino acids 112 to 117), but was identified at amino acid positions 1 to 28. Analysis of the Shannon entropy index for all codons in the 372 bp amino-terminal (N-terminal) region of the F protein of GenBank sequences (Supplementary Table 2; data not shown) identifies the amino acid positions 1–28 as a region of high entropy indicating heterogeneity at that location. The SNAP analysis, allowing graphic visualization of the distribution of the d*S* and d*N* values (Supplementary Table 2), confirmed the prevalence of negative selection across all NDV proteins with the exception of a few localized regions at the F (aa 1 to 28) and P proteins (aa 220 to 260).

Changes in virulence at the fusion protein cleavage site rarely occur during NDV evolution

In order to further investigate genomic evolution at the N-terminal end of the F protein and its relationship to the type of cleavage site present, phylogenetic analysis of the 372 bp region of the F protein that encoded the F2 peptide and the cleavage site was performed using a larger dataset. GenBank sequences from viruses isolated worldwide over the past 50 years and recent sequences from our lab were used to generate a dataset containing 861 virulent and 331 low



Class	Genotype	Virulent	Low-Virulence	Total
I (Total)	n/a	1*	187	188
II	I	2*	50	52
	II	19*	59	78
	IIa	0	34	34
	III	14	0	14
	IV	38	0	38
	V	158	0	158
	VI	343	0	343
	VII	251	0	251
	VIII	24	0	24
	IX	11	0	11
	n/a	1	0	1
II (Total)		861	143	1,004
Grand Total		**862**	**330**	**1,192**

Fig. 2. Distribution of virulent and low virulence viruses in NDV genotypes. The phylogenetic tree represents the genotype distribution of available NDV sequences corresponding to the 372 bp fusion gene fragment ($n = 1192$ non-redundant sequences). Genotypes are indicated in roman numerals for viruses from class II and virulence classification is based on the amino acid sequence at the fusion cleavage site. The numbers of viruses sorted by genotype and virulence are shown in the inset. Asterisks represent genotypes with both loNDV and vNDV.

virulence cleavage sites ($n = 1192$, Supplementary Table 5). A strong association between the type of cleavage site and viral genotype was detected (Fig. 2 and inset), suggesting that changes in virulence at the cleavage site rarely occur among viruses of the same genotype. Separation of low and high virulence viruses by genotype was also observed in the phylogenetic analysis of the partial *F* gene (Fig. 2). Genotypes III to IX represent highly virulent viruses, while genotypes I and II are predominately viruses of low virulence. In genotype I, only the viruses that caused the Australian outbreak were virulent, and in genotype II, a small subset of virulent viruses was isolated following the initial outbreaks in the 1940s in the United States.

Analysis of 1238 sequences corresponding to the N-terminal end of the F protein encoding the F2 peptide (positions 1 to 348) and the six critical amino acids of the cleavage site (positions 336 to 351) for adaptive evolution confirmed the occurrence of positive selection in a variable number of amino acids predominately at the N-terminal end of the F protein (Table 2). The number of positively selected sites varied slightly depending on how the viruses were grouped (by genotype or by virulence) (Supplementary Table 3). Negative or neutral selection pressure was detected at the F protein cleavage site, regardless of the dataset used (Table 2, and data not shown), suggesting that evolutionary pressures favor conservation of the cleavage site motif for viruses of low and high virulence.

Table 2
Detection of positive selection in the 372 bp sequences corresponding to the amino-terminal end of the fusion protein that encodes the F2 peptide and the cleavage site.

Selected dataset	Number of sequences	Overall dN/dS	Variance	Sites with dN/dS>1	*P*
Low virulence classes I and II	345	0.22	0.083	9	0.953
				13	0.967
				18	0.975
				22	0.926
				26	0.976
Virulent classes I and II	893	0.275	0.101	3	0.989
				4	0.999
				5	0.999
				9	0.994
				10	0.999
				11	0.999
				20	0.999
				28	0.999
All GenBank	1238	0.262	0.0837401	3	0.998
				4	0.999
				5	0.999
				9	0.999
				10	0.999
				11	1.000
				13	0.996
				18	0.924
				20	0.999
				28	0.999

Selected detail of class II d*N*/d*S* by genotype

Genotype	Number of sequences	d*N*/d*S*	Variance	Sites with d*N*/d*S*>1	*P*
II	78	0.27	0.116	5	0.973
				11	0.994
				17	0.969
V	108	0.25	0.182	10	0.996
VI	193	0.23	0.095	5	0.921
				6	0.994
				8	0.920
				19	0.948
				20	0.957
				28	0.998
				111	0.930
VII	200	0.358	0.227	4	0.999
				5	0.999
				11	0.998
				28	0.987
VIII	19	0.29	0.179	10	0.990
				13	0.983
				78	0.902

Dataset consisted of complete GenBank sequences, sorted by virulence of the cleavage site, or sorted by genotypes. Only those genotypes that had the largest number of sequences were included. Model 8 was used. Posterior cutoff for sites with d*N*/d*S*>1 = 0.9; *P* = posterior probability of sites with d*N*/d*S*>1.

Table 3
Difference in the clock rates between virulent and non-virulent viruses at the fusion coding sequences.

Clock rate GTR				
Summary statistic	Fusion protein			
	Low virulence		Virulent	
	Strict	Relaxed	Strict	Relaxed
Mean	2.28×10^{-4}	2.92×10^{-4}	1.32×10^{-3}	1.70×10^{-3}
SD of mean	1.22×10^{-6}	3.84×10^{-6}	2.74×10^{-6}	8.69×10^{-6}
Median	2.26×10^{-4}	2.83×10^{-4}	1.32×10^{-3}	1.68×10^{-3}
95% HPD lower	1.29×10^{-4}	1.19×10^{-4}	1.07×10^{-3}	1.22×10^{-3}
95% HPD upper	3.30×10^{-4}	4.79×10^{-4}	1.57×10^{-3}	2.21×10^{-3}
Effective sample size	1768	598	2189	861
	Complete coding sequences with intergenic regions			
Mean	5.04×10^{-4}	5.20×10^{-4}	1.25×10^{-3}	1.04×10^{-3}
SD of mean	6.74×10^{-7}	1.86×10^{-6}	1.27×10^{-6}	6.11×10^{-5}
Median	5.04×10^{-5}	5.17×10^{-4}	1.25×10^{-3}	9.21×10^{-4}
95% HPD lower	3.96×10^{-4}	3.76×10^{-4}	1.13×10^{-3}	2.70×10^{-4}
95% HPD upper	6.18×10^{-4}	6.75×10^{-4}	1.37×10^{-3}	2.01×10^{-3}
Effective	7192	1787	2471	80

The GTR substitution model with both Strict and Relaxed Lognormal clocks was evaluated using BEAST 1.4.6.

The fusion protein evolves at a faster rate in virulent viruses

The yearly rate of change was analyzed using a dataset corresponding to complete F proteins for which the year of isolation was available. Clock rate was calculated through Bayesian analysis as implemented in the BEAST program using the HKY and GTR models of nucleotide substitutions (Drummond and Rambaut, 2007). The BEAST program subjects the molecular sequence data to Bayesian Markov Chain Monte Carlo (MCMC) analysis and the data are oriented towards rooted, time-measured phylogenies inferred using strict and relaxed molecular clock models. Output is then analyzed to produce estimates of the parameters of interest, in this case, evolutionary rates. Analyses, verified by the Tracer and MCMC programs, were run (1×10^7 or 1×10^8) until the estimated samples sized (ESS) were large enough to produce reliable ESS (the chain length divided by the auto correlation time) (Rambaut and Drummond, 2007). The GTR substitution model allowing empirical bases frequencies, four categories of gamma distributed rates and a proportion of invariant sites was selected as the most fit. Both strict and relaxed lognormal molecular clocks were used.

The molecular clock rate for F proteins with a low virulence cleavage site motif ($n = 42$) was 2.28×10^{-4} (strict) and 2.92×10^{-4} (relaxed) (standard deviation of the mean (SD) $= 1.22 \times 10^{-6}$ and 3.84×10^{-6} respectively), as compared to virulent viruses for which the clock rate was 1.32×10^{-3} (strict) and 1.70×10^{-3} (relaxed) ($n = 76$) (Table 3). In addition, HKY evaluation for the F proteins for both virulent viruses and viruses of low virulence gave similar results (data not shown). The remarkable difference in the rate of change between these two groups of viruses prompted us to analyze genomic changes in viruses evolving under natural conditions (no vaccination).

We obtained viruses of low virulence from waterfowl and virulent viruses from cormorants isolated in the US during a similar period of time (Supplementary Table 4). These viruses are naturally endemic in wild birds and do not suffer the selective pressures of vaccination or human intervention. Genomes, except for the terminal 50–100 bp, of sixteen Class II viruses were sequenced and analyzed. Of these, eight were viruses of low virulence from genotype IIa isolated from waterfowl (1986 to 2004) and the other eight were genomes of virulent viruses from genotype V isolated from double crested cormorants (1992 to 2005). Genomes were sequenced using the random sequencing approach as developed in our laboratory (Kim, Suarez, and Afonso, 2006), and the rate of changes analyzed as above using the BEAST software package. Here we found the rates of 5.04×10^{-4} (strict) and 5.20×10^{-4} (relaxed) (SD of 6.74×10^{-7} to 1.86×10^{-6}, respectively) for the genomes of low virulence while for virulent genomes the rate was 1.25×10^{-3} (strict) and 1.04×10^{-3} (relaxed) (SD of 1.27×10^{-6} and 6.11×10^{-5}, respectively) (Table 3). Thus, these differences in genomic changes observed within viruses of the same genotype that were transmitted under natural conditions, suggest that the evolutionary rates of change may accelerate for viruses containing a virulent phenotype.

Discussion

We have analyzed the role of evolutionary forces on NDV genomes. Newcastle disease viruses encode their own RNA polymerases, produce long-lasting infections in wild birds, and induce cell fusion; conditions that favor recombination and rapid mutation rates. Previous studies reporting recombination in NDV sequences failed to demonstrate recombination as an evolutionary force, as viable recombinant progenies in nature had not been identified (Han et al., 2008; Qin et al., 2008a). Our analyses confirm the occurrence of events compatible with the action of RNA recombination in all of the other genes, except for the polymerase. However, only one breakpoint event that may represent a true evolutionary event was identified in the F protein coding regions of recent viruses from genotype V. In this particular case, the independent isolation and sequencing of multiple viruses by different laboratories, reduces the possibility of artificial creation of recombinant sequences (Fig. 1A). Although the true capacity of NDV to recombine can only be experimentally confirmed under highly controlled conditions, identification of this potential recombination event suggests that further work is warranted to investigate the role of recombination on NDV evolution.

Powerful maximum likelihood methods have now been developed that can detect evolutionary pressures upon individual genes or codons by comparing the d*N*/d*S* rates across the branches of a tree, including analysis of each codon and/or each branch separately (Yang, 2000; Yang et al., 2000) (Yang and Nielsen, 2002). This provides the opportunity to examine the previously unexplored role of selection pressures on the evolution of NDV. Using robust statistical methods, both positive and negative selection was identified in various NDV proteins. Overall negative or neutral selective pressures were detected in each of the coding regions, with codon-specific positive selection observed in five NDV proteins. The removal of putative recombinants reduced the number of positively selected amino acids and emphasized the need for caution in the interpretation of the data from crude datasets. For example, data in Table 1 indicates that even the most robust methods (likelihood, M8 under Bayesian selection) of detection and selection can be affected if the levels of recombination in the dataset are elevated (Anisimova, Nielsen, and Yang, 2003). Of significant interest is the presence of negative selection across the F protein cleavage site, regardless of the virulence or genotype, suggesting that conservation of the fusion cleavage site motif is likely important for the persistence of NDV in nature. Interestingly, the region surrounding the cleavage site was also under negative pressure (data not shown). This region in NDV, as well as in other paramyxovirus F proteins, is part of an alpha helix that extends from amino acids 76 to 105 and is critical to the proper folding of the molecule. Mutational analysis in the measles virus F protein has shown that mutations in this domain affect syncytia formation (Morrison, 2003; Plemper and Compans, 2003), and the putative third helical region with heptad repeats (HR-C) region (residues 102 and 109) of F2 is also thought to be required for folding and fusogenic activity of the F protein (Plemper and Compans, 2003).

The potential of highly virulent viruses emerging from low virulence strains is cause for concern for poultry producers worldwide. There is evidence that circulating low virulence viruses may have mutated to cause outbreaks in the Republic of Ireland (1990) and Australia (1998 through 2000). Experimental studies demonstrating the capacity of low virulence viruses to increase in virulence with passage in chickens, highlight this concern (Islam et al., 1994). In the Australian outbreak, only two nucleotide changes at the F protein cleavage site were necessary to change it from a trypsin-like (low virulence) to a furin-like (virulent) site. Results from our study suggest that the F protein cleavage site of the ancestor viruses tends to be conserved and that exceptional circumstances may be required to allow the mutation from a low to highly virulent cleavage site (Table 2 and Fig. 2). The stability of vaccine viruses, such as LaSota and B1 that have been used as live bird vaccines for over 40 years, along with the lack of evidence of a virulent virus reverting to low virulence among the sequences analyzed (data not shown) also supports this hypothesis. The prevalence of negative selection on the F protein of NDV (Supplementary Table 2), in addition to the lower rate of change (mutations per year) observed in low virulence viruses (Table 3) suggests that it may be difficult for a virulent virus to emerge from one of low virulence. Although the possibility of creating compensatory mutation at other proteins or sites within the *F* gene cannot be disregarded, the existence of negative selection across the majority of the NDV coding sequences suggests that these changes do not happen frequently. The frequent occurrence of positive selection at a ~28 codon region near the N-terminal end is not surprising considering that this region is highly variable and that positive selection is likely to operate in regions of a protein where a high level of structural diversity is not required. For example, positive selection has been detected on surface residues of the gp120 envelope gene of the human immunodeficiency virus (HIV-1) and the hemagglutinin (HA) of the influenza virus (Yang et al., 2000, 2003).

The presence of higher rates of evolution in virulent genomes obtained from wild birds or in the fusion protein of random selected samples (including a large percentage of sample of poultry origin) (Table 3) suggest that the phenomena may be general in nature and perhaps confer virulent viruses additional evolutionary advantages that allow persistence in nature. Low virulence viruses have the advantage of allowing the host to live for longer periods of time, thus increasing the chances for viral replication. Virulent viruses, with a theoretically reduced time for replication, have nonetheless continued to evolve in nature (Perozo et al., 2008; Qin et al., 2008b). Since virulent viruses are capable of infecting a broader range of host tissues, one possibility for their enhanced evolution is that the increased number of target cells infected by virulent viruses may contribute to expand the diversity of the viral population, and therefore improve the capacity of the virus to adapt and evolve.

Materials and methods

Viruses, RNA isolation, and RT-PCR

Twenty-seven Newcastle disease isolates in allantoic fluid were obtained from the SEPRL and USGS NDV repository. Ribonucleic acid (RNA) was extracted using Trizol LS (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. After extraction, RNA was eluted in 100 µl of RNase-free water and stored at −80 °C. PCR amplification of the RNA was performed using the Qiagen One-Step RT-PCR kit (Qiagen, Valencia, CA). Amplified products were separated on a 1% agarose gel, bands were excised and eluted using the QIAquick Gel extraction kit (Qiagen) and the samples were quantified using a standard spectrophotometer.

Sequence data, nucleotide sequencing, alignment analysis, and entropy data

All double-stranded nucleotide sequencing reactions were performed with fluorescent dideoxynucleotide terminators in an automated Applied Biosystems International (ABI) sequencer as previously reported (Kim et al., 2007). Random genome nucleotide sequencing was done as reported previously for Avian Influenza (Kim et al., 2006). Nucleotide sequence assembly and editing were conducted with the LaserGene sequence analysis software package or with Codon Aligner using Phred and Phrap for complete genomes (http://www.codoncode.com/). Forty-three novel sequences were submitted to GenBank and the accession numbers are FJ705452 through FJ705478 and GQ288377 through GQ288392. Sequences retrieved from GenBank public databases were used to generate alignments; accession numbers can be obtained as Supplementary material. Alignments of non-redundant complete coding regions with and without recombination (n = total/non-recombinant only) were used for determination of positive selection in the following open reading frames: NC (n = 82/72), P (n = 91/75), M (n = 74/70), F (n = 129/115), HN (n = 122/119), and L (n = 43/43). A dataset containing GenBank sequences encoding the 374-nucleotide *F2* gene fragment (n = 1238) was used for Table 2. Of these 1238 sequences, 1192 could be classified into defined genotypes and used to create Fig. 2. Datasets can be viewed in Supplementary Table 5. Alignments were performed using BioEdit v. 5.0.9 with either the ClustalW (Thompson et al., 1994) program or the Muscle (Edgar, 2004) program, followed by manual editing using Molecular Evolutionary Genetics Analysis (MEGA) (Kumar et al., 2004). The *V* gene that originates from a frame shift on the P coding region was not included in this study due to insufficient availability of sequences. Sequences from the first through last codons were aligned and did not include indels or termination codons. The variability at each position in the *F* gene coding region for these NDV isolates was calculated by using the entropy algorithm available from the BioEdit software (Hall, 1999).

Analysis of recombination

To characterize the presence of recombination in NDV isolates, complete coding regions obtained through sequencing and from available GenBank datasets were analyzed for recombination using the split decomposition method and the following local statistical methods: Gene Conversion (Geneconv) (Sawyer, 1989), Maximum chi-square test (MaxChi) (Maynard Smith, 1992), maximum mismatch chi-square (Chimaera) (Posada and Crandall, 2001), Bootscan (Salminen et al., 1995), sister-scanning (SiScan) (Gibbs, et al., 2000), and 3Seq (Boni et al., 2007) as implemented in the recombination detection program (RDP) version 3.24 (Martin et al., 2005) for each of the NDV genes. RDP3.24 is available from http://darwin.uvigo.es/rdp/rdp.html for the Windows OS and SplitsTree is available at www.splitstree.org (Huson and Bryant, 2006). For RDP2.34 all sequences were considered to be linear and the *P*-value cutoff was set to 0.01. The standard Bonferroni correction, consensus daughters finding, and breakpoints polishing methods were also used for analysis with the RDP2.34 program. For Split trees, the general time reversible (GTR) model of nucleotide substitution was used on the split decomposition program. GTR refers to the relative substitution rates for A↔C, A↔G, A↔T, C↔G and G↔T in this model of nucleotide substitution (Rodriguez et al., 1990). In order to identify true recombination events, the following criteria were followed: 1) identification of an event using at least four independent detection methods ($P<0.01$); 2) isolation and/or sequencing of the viruses independently by different authors; and 3) identification of a natural lineage of descendant viruses. Accession numbers for sequences from Supplementary Table 1 was used to create Fig. 1.

Adaptive evolution and synonymous/non-synonymous analysis of NDV nucleotide sequences

Preliminary detection of selection pressures at codon sites was completed with the maximum likelihood method implemented in CODEML of the PAML 4.0 package in a comparison of different codon-based models that allow for variable selection among the sites (Yang, 2000; Yang and Bielawski, 2000; Yang and Nielsen, 2002; Yang et al., 2000; Yang and Swanson, 2002). We compared models MO, M2, M3, M7 and M8 using the likelihood ratio test and selected M8 (Beta + ω) as implemented by HYPHY as the model for this study (Pond, Frost, and Muse, 2005). Bayes Empirical Bayes probabilities (BEB) approach was employed to identify specific residues under positive selection.

The distribution of d*S* and d*N* values across the length of each protein was graphically visualized using the cumulative output values from the synonymous/non-synonymous analysis program (SNAP) available at http://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html. This program calculates d*S* and d*N* substitution rates on a set of full length codon-aligned nucleotide sequences based on the method of Nei and Gojobori (1986), incorporating a statistic developed by Ota and Nei (1994). CodeML Model 8 (beta) implemented in HYPHY (http://www.hyphy.org) was used for d*N*/d*S* ratios and positively selecting sites listed in Table 1. Data and accession numbers for sequences used can be found in Supplementary Table 2.

Evolutionary analysis of NDV proteins

Phylogenetic tree construction (Figs. 1A and 3) was performed with Phylogenies by Maximum Likelihood (PhyML) V2.4.4 under the GTR model of nucleotide substitution with estimated proportions of invariable sites, maximum likelihood (ML) base frequencies estimates, four substitution rate categories, and an optimized gamma distribution parameter (Guindon and Gascuel, 2003). The best fit substitution model (GTR + I + G) for the data was determined by Akaike's information criterion (AIC), a measure of how well a model explains the data, in Modeltest 3.6 using Phylogenetic Analysis Using Parsimony (PAUP) (Posada and Crandall, 1998). Sequences can be found in Supplementary Tables 1 and 5.

Rate of change of NDV proteins

The annual rate of change (clock rate) for low virulence and virulent viruses was calculated using Bayesian Evolutionary Analysis Sampling Trees (BEAST 1.4.6) (Drummond and Rambaut, 2007) under the GTR and Hasegawa–Kishino–Yano (HKY) models for the coding regions corresponding to the full F and NC proteins (Hasegawa et al., 1985). The F protein sequences were compared in addition to the complete coding sequences and intergenic regions of both virulent and low virulence viruses (Table 3). The Markcov Chain Monte Carlo (MCMC) analysis were run for 1×10^7 or 1×10^8 for the relaxed lognormal complete genome (second genome analysis) datasets which states samples every 1000 generations and the initial 10% burn in samples are discarded. The GTR substitution model allowing empirical bases frequencies, four categories of gamma distributed rates and a proportion of invariant sites were selected as the most fit for the data. A constant coalescent size, with Jeffrey priors on either a strict or a relaxed lognormal clock, was used. The adequacy of sampling was assessed via effective sample size (ESS), which was larger than 500, except for the virulent complete genome dataset (n = 80). Sequences can be found in Supplementary Table 4.

Acknowledgments

We gratefully acknowledge Kevin Winker, Michael Day, and David Suarez for the critical review of this manuscript; Dawn Williams-Coplin for the technical assistance; Alexei Drummond for the support on the use of BEAST, and the South Atlantic Area Sequencing Facility for nucleotide sequencing.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.virol.2009.05.033.

References

Afonso, C.L., 2008. Not so fast on recombination analysis of *Newcastle disease virus*. J. Virol. 82 (18), 9303.

Alexander, D.J., Campbell, G., Manvell, R.J., Collins, M.S., Parsons, G., McNulty, M.S., 1992. Characterisation of an antigenically unusual virus responsible for two outbreaks of Newcastle disease in the Republic of Ireland in 1990. Vet. Rec. 130 (4), 65–68.

Alexander, D.J., Gough, R.E., Saif, Y.M., Barnes, H.J., Glisson, J.R., Fadly, A.M., McDougald, L.R., Swayne, D.E., 2003. Newcastle disease, other avian Paramyxoviruses, and Pneumovirus infections. Disease of Poultry, vol. 11th. Iowa State Press, Ames, IA, pp. 63–92.

Anisimova, M., Nielsen, R., Yang, Z., 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164 (3), 1229–1236.

Boni, M.F., Posada, D., Feldman, M.W., 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics 176 (2), 1035–1047.

Bush, R.M., (2001). Predicting adaptive evolution 2(5), 387–392.

Chare, E.R., Gould, E.A., Holmes, E.C., 2003. Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. J. Gen. Virol. 84 (Pt 10), 2691–2703.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.

Duarte, E.A., Novella, I.S., Weaver, S.C., Domingo, E., Wain-Hobson, S., Clarke, D.K., Moya, A., Elena, S.F., de la Torre, J.C., Holland, J.J., 1994. RNA virus quasispecies: significance for viral disease and epidemiology. Infect. Agents. Dis. 3 (4), 201–214.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nuclei. Acids Res. 32 (5), 1792–1797.

Gibbs, M.J., Armstrong, J.S., Gibbs, A.J., 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16 (7), 573–582.

Glickman, R.L., Syddall, R.J., Iorio, R.M., Sheehan, J.P., Bratt, M.A., 1988. Quantitative basic residue requirements in the cleavage-activation site of the fusion glycoprotein as a determinant of virulence for *Newcastle disease virus*. J. Virol. 62 (1), 354–356.

Gould, A.R., Kattenbelt, J.A., Selleck, P., Hansson, E., la-Porta, A., Westbury, H.A., 2001. Virulent Newcastle disease in Australia: molecular epidemiological analysis of viruses isolated prior to and during the outbreaks of 1998–2000. Virus Res. 77 (1), 51–60.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52 (5), 696–704.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. 41, 95–98.

Han, G.Z., He, C.Q., Ding, N.Z., Ma, L.Y., 2008. Identification of a natural multi-recombinant of *Newcastle disease virus*. Virology 371 (1), 54–60.

Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22 (2), 160–174.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23 (2), 254–267.

Islam, M.A., Ito, T., Takakuwa, H., Takada, A., Itakura, C., Kida, H., 1994. Acquisition of pathogenicity of a *Newcastle disease virus* isolated from a Japanese quail by intracerebral passage in chickens. Jpn. J. Vet. Res 42 (3–4), 147–156.

Kim, L.M., Suarez, D.L., Afonso, C.L., 2006. Biotechnologia e Saude Animal. Viscosa, Brazil.

Kim, L.M., King, D.J., Curry, P.E., Suarez, D.L., Swayne, D.E., Stallknecht, D.E., Slemons, R.D., Pedersen, J.C., Senne, D.A., Winker, K., Afonso, C.L., 2007. phylogenetic diversity among low virulence Newcastle disease viruses from waterfowl and shorebirds and comparison of genotype distributions to poultry-origin isolates. J. Virol. 81 (22), 12641–12653.

Kosiol, C., Bofkin, L., Whelan, S., 2006. Phylogenetics by likelihood: evolutionary modeling as a tool for understanding the genome. J. Biomed. Informatics 39 (1), 51–61.

Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. 5 (2), 150–163.

Lai, M.M., 1992. RNA recombination in animal and plant viruses. Microbiol. Rev. 56 (1), 61–79.

Martin, D.P., Williamson, C., Posada, D., 2005. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics 21 (2), 260–262.

Maynard Smith, J., 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. 35, 126–129.

Morrison, T.G., 2003. Structure and function of a paramyxovirus fusion protein. Biochim. Biophys. Acta 1614 (1), 73–84.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3 (5), 418–426.

OIE, B.S.C., 2004. Manual of Diagnostic Tests and Vaccines for Terrestrial Animals: Mammals, Birds and Bees, 5 ed. Office international des âepizooties, Paris. Chapter 2.1.15. 2 vols.

Ota, T., Nei, M., 1994. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. Mol. Biol. Evol. 11 (4), 613–619.

Perozo, F., Merino, R., Afonso, C.L., Villegas, P., Calderon, N., 2008. Biological and phylogenetic characterization of virulent *Newcastle disease virus* circulating in Mexico. Avian. Dis. 52 (3), 472–479.

Plemper, R.K., Compans, R.W., 2003. Mutations in the putative HR-C region of the measles virus F2 glycoprotein modulate syncytium formation. J. Virol. 77 (7), 4181–4190.

Pond, S.L., Frost, S.D., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21 (5), 676–679.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14 (9), 817–818.

Posada, D., Crandall, K.A., 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. PNAS 98, 13757–13762.

Qin, Z., Sun, L., Ma, B., Cui, Z., Zhu, Y., Kitamura, Y., Liu, W., 2008a. F gene recombination between genotype II and VII *Newcastle disease virus*. Virus Res. 131 (2), 299–303.

Qin, Z.M., Tan, L.T., Xu, H.Y., Ma, B.C., Wang, Y.L., Yuan, X.Y., Liu, W.J., 2008b. Pathotypical characterization and molecular epidemiology of *Newcastle disease virus* isolates from different hosts in China from 1996 to 2005. J. Clin. Microbiol. 46 (2), 601–611.

Rambaut, A., Drummond, A.J., 2007. Tracer. http://beast.bio.ed.ac.uk/Tracer.

Rodriguez, F., Oliver, J.F., Marin, A., Medina, J.R., 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. 142, 485–501.

Rott, R., 1979. Molecular basis of infectivity and pathogenicity of Myxovirus. Arch. Virol. 59, 285–298.

Salminen, M.O., Carr, J.K., Burke, D.S., McCutchan, F.E., 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res. Hum. Retroviruses 11, 1423–1425.

Sawyer, S.A., 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. 6, 526–538.

Spann, K.M., Collins, P.L., Teng, M.N., 2003. Genetic recombination during coinfection of two mutants of human respiratory syncytial virus. J. Virol. 77 (20), 11201–11211.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (22), 4673–4680.

Worobey, M., Holmes, E.C., 1999. Evolutionary aspects of recombination in RNA viruses. J. Gen. Virol. 80 (Pt 10), 2535–2543.

Yang, Z., 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J. Mol. Evol. 51 (5), 423–432.

Yang, Z., Bielawski, J.P., 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. 15 (12), 496–503.

Yang, Z., Nielsen, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19 (6), 908–917.

Yang, Z., Swanson, W.J., 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol. Biol. Evol. 19 (1), 49–57.

Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.M., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155 (1), 431–449.

Yang, W., Bielawski, J.P., Yang, Z., 2003. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J. Mol. Evol. 57 (2), 212–221.

Glossary

NDV: *Newcastle disease virus*
APMV-1: Avian paramyxovirus type 1 virus
ND: Newcastle disease
F: fusion
HN: hemagglutinin-neuraminidase
P: phosphoprotein
M: matrix
L: RNA polymerase
N: nucleocapsid
R: arginine
K: lysine
MEGA: Molecular Evolutionary Genetics Analysis
GTR: general time reversible
PhyML: Phylogenies by Maximum Likelihood
ML: maximum likelihood
AIC: Akaike's information criterion
PAUP: Phylogenetic Analysis Using Parsimony
PAML: Phylogenetics Analysis Using Maximum Likelihood
HYPHY: Hypothesis Testing Using Phylogenies
BEB: Bayes Empirical Bayes probabilities
BEAST: Bayesian Evolutionary Analysis Sampling Trees
HKY: Hasegawa–Kishino–Yano
dN: non-synonymous substitutions per synonymous site
dS: synonymous substitutions per synonymous site
ω: d*N* to d*S* ratio
WFL: wild-water fowl